AD-A054 952 POLYTECHNIC INST OF NEW YORK BROOKLYN F/G 17/2
STUDY AND DEVELOPMENT OF SPEECH-SEPARATION TECHNIQUES.(U)
MAY 78 T W PARSONS F30602-77-C-0013

UNCLASSIFIED RADC-TR-78-105 NL

| OF |

AD
A054952

END
DATE
FILMED
7-78
DDC

1 B.S.

RADC-TR-78-105
Final Technical Report
May 1978

STUDY AND DEVELOPMENT OF SPEECH SEPARATION TECHNIQUES

Thomas W. Parsons

Polytechnic Institute of New York

Approved for public release; distribution unlimited.

ROME AIR DEVELOPMENT CENTER
Air Force Systems Command
Griffiss Air Force Base, New York 13441

This report has been reviewed by the RADC Information Office (OI) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

RADC-TR-78-105 has been reviewed and is approved for publication.

APPROVED:

ROBERT A. CURTIS, Captain, USAF
Project Engineer

APPROVED: H Davis

HOWARD DAVIS
Technical Director
Intelligence & Reconnaissance Division

FOR THE COMMANDER: John P. Huss

JOHN P. HUSS
Acting Chief, Plans Office

If your address has changed or if you wish to be removed from the RADC mailing list, or if the addressee is no longer employed by your organization, please notify RADC (IRAA) Griffiss AFB NY 13441. This will assist us in maintaining a current mailing list.

Do not return this copy. Retain or destroy.

19

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
18 RADC-TR-78-105

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
6 STUDY AND DEVELOPMENT OF SPEECH-SEPARATION TECHNIQUES.

A038637

5. TYPE OF REPORT & PERIOD COVERED
9 Final Technical Report.

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
10 Thomas W. Parsons

8. CONTRACT OR GRANT NUMBER(s)
15 F30602-77-C-0013

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Polytechnic Institute of New York ✓
333 Jay Street
Brooklyn NY 11201

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
31011G
70550732

11. CONTROLLING OFFICE NAME AND ADDRESS
Rome Air Development Center (IRAA)
Griffiss AFB NY 13441

12. REPORT DATE
11 May 1978

13. NUMBER OF PAGES
42
12 54p.

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
Same

15. SECURITY CLASS. (of this report)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
N/A

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Same

18. SUPPLEMENTARY NOTES

RADC Project Engineer: Captain Robert A. Curtis (IRRA)

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Two-talker separation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A frequent problem in speech communication is that of restoring the intelligibility of speech which has suffered from interference. The interference may consist of noise, periodic signals, or competing speech. Suppressing interference caused by competing speech is a particularly difficult problem, because there are no simple discriminants which can be used to distinguish speech from speech.

In a previous contract (F30602-74-C-0175), a speech-separation technique was

408 704

developed which worked by inspecting the Fourier transform of the combined speech and selecting the harmonics of the desired voice. This process was used to separate messages whose content was restricted to vowels and vowel-like sounds (e.g., "We were away a year ago.") and whose spectra therefore exhibited the periodicity on which the technique depends. Experiments with speech containing non-vocalic sounds, however, suggested that the process could probably be made to separate such speech as well.

In the present research effort, the separation process has been extended to handle natural speech--i.e., speech not restricted to vowel sounds. Natural speech presents many other problems for the separation process besides the presence of non-vocalic sounds. The harmonic components of vocalics are less clear and well-defined when they occur in natural speech, and pitch fluctuates much more widely and rapidly than in vocalic speech. Therefore all phases of the separation process, and especially the pitch-extraction routine, had to be made robust enough to work reliably in a natural-speech environment. A related problem was to enable the process to identify and characterize the desired harmonics accurately enough that the output speech will sound clear and natural. To make these improvements in the process, its performance on typical test data was studied and causes of failure identified; then modifications were devised to prevent these failures. Finally, the process must be able to handle non-vocalic sounds. In human listeners, perception of non-vocalics depends heavily on cues contained in the adjacent vocalic sounds, and if these cues are preserved, then in most cases the ear will perceive the desired sound. It was found that the separation process preserves these cues well and that most non-vocalics, with the exception of some initial unvoiced sounds, are clearly audible in the output speech.

The revised version of the process has been tested with samples of speech recorded from commercial radio broadcasts. The results show greatly improved performance in almost every respect, although the quality of the recovered speech still needs further improvement. Although the process is still not error-free, the remaining errors are infrequent enough that it is practical to remove them by user intervention. Future research should be devoted to further improvements in the process, to the possible extension of the process to more than two talkers and to noisy data, and to the consideration of an interactive processing system.

ABSTRACT

One of the most common types of interference in speech communication is that caused by the speech of a competing talker. A technique has been developed for suppressing such interference by examining the Fourier transform of the input and selecting the harmonics of the desired voice. The initial version of this process was applicable only to vocalic speech (i.e., speech consisting only of vowels and vowel-like sounds), but in subsequent research steps have been taken to extend the process to natural (i.e., unrestricted) speech. This report describes the improvements which have been made in this research, first, to ruggedize the process so that it can perform in an natural-speech environment, second, to improve the intelligibility and naturalness of the recovered speech, and third, to enable the process to handle the non-vocalic speech sounds (such as plosives and fricatives) which occur in natural speech.

CONTENTS

CONTENTS (Continued)

EVALUATION

The capability to separate two talkers talking simultaneously on the same channel represents a serious and recurring communication problem. This report summarizes current efforts and the techniques developed to date in order to separate the speech of two simultaneous talkers.

While the current separation process is intelligible, the output is not of high quality and additional research is necessary to improve the synthesis procedure so that both the output speech quality and intelligibility are improved.

ROBERT A. CURTIS, Capt, USAF
Project Engineer

1.0 INTRODUCTION

A recurrent problem in speech communication is that of clarifying speech whose intelligibility has been degraded by interference. Of the various types of interference likely to occur, one of the commonest, and most difficult to suppress, is the speech of a competing talker. In a previous contract (F30602-74-C-0175), a preliminary investigation of the problem of separating the voices of two talkers was carried out. In the course of that research, a technique was developed which separated vocalic (i.e., vowel or vowel-like) speech sounds by inspecting the Fourier transform of the speech and identifying the harmonics of the desired voice. This technique was successfully used to separate (a) vowel sounds sung at constant and varying pitch, (b) vocalic sentences ("Were you away a year ago?") sung at constant and varying pitch, and (c) vocalic sentences spoken with natural intonation. Attempts to apply the process to intoned natural (i.e., not all-vocalic) speech yielded poor-quality output, although the results suggested that this was not due to any inherent limitation of the process. In addition, the fact that many non-vocalics seemed to be present in the output suggested that the process might be extended to fully-natural speech more readily than might have been expected.

In the present research effort, an attempt has been made to extend the process to natural speech. To do this, we have identified and attempted to correct the weaknesses in the process which degraded its performance in the initial experiment with natural speech. The revised version of the separation process has been tested with samples of speech recorded from commercial radio broadcasts. The results show greatly improved performance in every respect, although the quality of the recovered speech still needs further improvement.

1.1 Summary of Existing Process

In this report, we assume familiarity with the existing separation process as described in RADC-TR-75-155, "Enhancing Intelligibility of Speech in Noisy or Multi-Talker Environments", but it will be useful to review it briefly here. The steps in the process are as follows:

1. The digitized input signal is divided into overlapping segments; the segments are windowed (i.e., time-weighted) and Fourier transformed.
2. Spectrum peaks (i.e., maxima in the modulus of the Fourier transform) are identified and peak tables compiled. These tables give the parameters (frequency, amplitude, and phase) of each spectrum peak.
3. Peak overlaps, in which the k^{th} harmonic of one talker's pitch is nearly equal to the n^{th} harmonic of the other talker, are detected and their components separated. The parameters of the components replace those of the composite in the peak tables.
4. Each talker's pitch is determined by an examination of the spectrum peaks as listed in the peak tables. Each talker's pitch harmonics are then identified by scanning the peak tables.
5. Consistency of talker identification is obtained by tracking the two talkers' pitches, using a pair of predictive filters and a simple rule for matching the current pitches to the predictions.
6. Each individual's speech is recovered by synthesizing a new

Fourier transform containing only those spectrum peaks which have been assigned to him.

7. The output speech is obtained by inverse-transforming the result and adding the overlapping time-segments together to form a continuous signal.

1.2 Areas Selected for Study

When the process described above was tried with natural speech, the following problems were encountered:

a. Errors in pitch detection were much more frequent than with vocalic speech--so numerous, in fact, that the process could not be applied without forcing the pitch to the correct value.

b. The unwanted talker's voice was not completely suppressed; there remained a considerable amount of crosstalk, particularly in the voice having the lower amplitude.

c. The recovered speech sounded muffled and indistinct, as though it had been low-pass filtered.

d. The pitch tracker could not follow the abrupt changes in pitch which occurred and could not acquire rapidly-changing pitch tracks.

These various problems were all traceable to various deficiencies in the corresponding steps in the process, but it was also clear that every phase of the separation process would have to be made more robust in order to cope with the environment created by natural speech. On the basis of these observations, the following areas were chosen for study and improvement in the present project.

1.2.1 Peak Separation

It was clear that not all spectrum-peak overlaps were being properly separated. It was found on inspection that a certain number of close overlaps were in fact being resolved incorrectly--that is, the frequencies of the estimated components were inconsistent with the values which they had to have in order to fit into their respective harmonic trains. These inaccurate frequencies then provided faulty data to the pitch program and degraded its performance; they were also subsequently used by the output synthesizer and in all probability also degraded the quality of the output speech. In addition, the existing peak-separation method was crude and oversimplified: it assumed that parameters of one of the components were already known with reasonable accuracy, and in many cases this was not so; furthermore, there was no satisfactory way to account for the effects of changing pitch on the peak shapes (FM effects).

1.2.2 Pitch Extraction

The pitch-determining procedure was probably the weakest link in the process as applied to natural speech; it had been necessary to force it to the correct pitch values in order to process monotone natural speech. Performance with vocalic speech had been good, but partly because the input signal was well-behaved (i.e., of nearly constant amplitude and with slowly-varying pitch) and partly because the routine received guidance from the pitch tracker. Conditions which made this guidance most urgently needed were also those which were most likely to make the tracker fail, and in any case, the performance of the tracker with natural speech was not good enough to be used as a guide.

1.2.3 Peak Assignment

Peak assignment is the name given to the process which identifies each talker's pitch harmonics in the peak tables after pitch has been determined. The poor quality of natural speech seemed attributable principally to errors in this process. Crosstalk could be traced to errors assigning to one talker harmonic peaks that actually belonged to the other; the muffled sound resulted from inability to find the correct harmonics above a certain frequency--in many cases, above the first large gap between formants. The existing assignment rule functioned like a fading-memory filter, but it was apparently both too rigid to adapt well to cumulative errors and at the same time not rigid enough to reject spurious peaks.

1.2.4 Tracker Acquisition and Prediction

The tracker performed very well with vocalic utterances, but this was partly because the data were so easy to handle. The tracks it had to follow were long, smooth (i.e., with limited second derivative), unbroken, and easy to acquire (i.e., with limited first derivative). With natural speech, the tracks tend to be short, frequently interrupted by plosives and other unvoiced sounds, irregular, and occasionally steep. Hence a tracker was needed which could acquire steep tracks easily, could follow (and predict) irregular tracks with greater accuracy, and could bridge gaps in the tracks.

1.2.5 Handling Non-vocalic Speech Sounds

The separation method as originally developed was intended for use only with vocalic sounds (i.e., vowels, glides, liquids, nasals, and possibly some voiced stops). The experiment with intoned natural speech showed that many nonvocalics were also perceived clearly in the recovered speech. It is believed that this was due to cues furnished by formant transitions, but it was by no means certain how far this effect could be relied on, and it was felt that some special treatment would be required for nonvocalic sounds. It was necessary to find ways of detecting these sounds as well as ways of reproducing them in the output speech.

2.0 PEAK SEPARATION

The peak-separation process developed in the initial research phase worked as follows: it was assumed that the larger component essentially determined the shape of the overlap, and hence that the parameters (frequency, amplitude, phase) of the larger component could be estimated from those of the composite. Knowing from theoretical considerations (as explained in RADC-TR-75-155, section 3.5) what the shape, W(x), of a component peak ought to be, we subtracted an estimate of the larger component from the overlap. The residue was then examined for peaks, and the largest peak in the residue was taken to be the smaller component. The principal defect of this method was that in close overlaps (i.e., those overlaps in which the two components were close in both amplitude and frequency), it was not possible to estimate the larger component's parameters correctly from those of the overlap. There will, of course, *always* be *some degree* of closeness, or some distance in parameter space, within which it will be impossible to resolve the components; this part of the research can be viewed as an attempt to minimize this distance.

In the present project, a measure of peak "quality" was developed for the purpose of evaluating separation performance, and two alternative peak-separation techniques were explored.

2.1 Peak Quality

The two-talker program includes routines for computing an ideal spectrum peak shape with any required set of parameters. Hence it is a simple matter, given an actual peak, to compare its shape with that of an ideal peak having the same parameters. Since the shape of each separated component should ideally match the theoretical shape (assuming uniform signal level over the time window), similarity to the theoretical shape seems a reasonable measure of peak quality. Accordingly, a sum-squared-error quality test was implemented as follows: if the complex value of the observed peak at the n^{th} sample point is S_n and the complex value of the ideal peak is F_n, then we define the error e_n as $S_n - F_n$, and the quality is given by

$$q = 1 - \sum_n {e_n}^2 \ / \sum_n {S_n}^2 \qquad (2\text{-}1)$$

where the sums are taken over the main lobe and the first two sidelobes of the ideal peak. This quality measure was used in both of the following separation methods as a way of monitoring the degree of success obtained. The measure ranges from upwards of .999 for isolated harmonics in quiet vocalic speech to approximately -.02 in badly distorted overlaps. The quality measure is now treated as another parameter of each spectrum peak, and the peak tables include an array giving the q-value of every spectrum peak.

2.2 Iterative Peak Separation

If the major component can be estimated at all well, then it must be possible to get at least a rough approximation of the smaller peak's parameters by inspecting the residue. In that case, if an ideal smaller peak is then subtracted from the composite, an improved estimate of the major peak's parameters should be available from this new residue. In that case, the process can be repeated and a closer approximation to the smaller peak obtained. In fact, the process can be iterated until there is no further significant improvement in the combined qualities of the isolated components. (The quality of the isolated component is found by computing the q-value of the residue from which we have estimated the component.) This leads to the following procedure:

a. Subtract estimate of major peak (i.e., ideal peak with estimated parameters of major peak).

b. Find quality and parameters of minor peak from residue.

c. Subtract ideal estimate of minor peak.

d. Find quality and parameters of major peak from residue.

e. If improvement in combined quality (as compared to previous iteration) is greater than 0.1%, return to Step a.

It will be seen that the process requires very little encouragement to iterate. The 0.1% figure was arrived at because larger thresholds caused the process to terminate before converging on what we knew to be correct values. (The process is made to stop after six iterations in any case, in order to keep execution-time requirements within reasonable bounds.)

At the outset of the project, we had estimated that the old peak separation process successfully separated approximately 75 per cent. of all overlaps; the iterative method successfully separates approximately 80 per cent. (An overlap is successfully separated if the components obtained fit into the two talkers' harmonic series.) Nevertheless, the performance of this process depends on the quality of the parameter estimates with which it starts. If the initial estimates are good, the process converges to a pair of peaks with q-values close to unity; if they are far off, however, there is no guarantee that the process will find its way to the correct values. Hence the iterative method, although it has been incorporated in the process in place of the old method, is not a solution to the really close overlaps from which it is impossible to find good starting estimates.

2.3 Parameter-Space Searching

The essential parameters of a harmonic component are frequency, amplitude, phase, and FM rate. These parameters may be thought of as a vector in a 4-dimensional parameter space, in which case the peak separation problem is one of finding a pair of vectors which account for the shape of the observed peak. The iterative method, just described, starts with an estimate of one of these vectors and attempts to make the composite tell it where the other vector should be. Another possibility is to make a search of the parameter space itself, starting with an initial pair of estimated vectors (which can be obtained from the basic method) and perturbing them so as to maximize the quality of the fit to the observed data. Such a search is, in general, an expensive undertaking, since it proceeds by trial and error and requires a lot of computer time per trial. Two search techniques were investigated in this project. Both techniques use the principle of varying one parameter at a time for one peak at a time. For each new combination of parameter values, a pair of ideal peaks is synthesized and their sum compared to the input data to obtain a measure of goodness of fit. The techniques differ only in the way the parameter under consideration is varied.

2.3.1 Interval-Halving Search

This is an adaptation of the well-known search method used in numerical analysis. The parameter to be varied is changed by a fixed step size, and for each new value, the quality of the combined peaks is computed and compared to that for the preceding value. When a maximum is passed, the step size is halved and its sign reversed. This procedure is repeated until there is no further significant improvement in the overall quality; then the process passes to the next parameter.

Although this method has the advantage that it always converges, it consumes large quantities of computer time. The ideal peak is computed as follows: let the time weighting function be w(t) and its Fourier transform be W(f). Then for a component with a center frequency f_o, the peak shape is approximately

$$F(x) = W(x) + jk\mu T^2 W''(x), \tag{2-2}$$

where $x = f - f_o$, μ is the FM rate, T is the duration of the time window, and k a constant of proportionality. Therefore, every time an ideal peak is computed, the program must compute W(x) and its second derivative for every point x. For Hanning weighting, used in the two-talker process, this means computing

$$W(x) = \sin\pi x/[\pi x(1 - x^2)],$$

$$W'(x) = [\cos\pi x + (3x^2 - 1)W(x)]/[x(1 - x^2)],$$

and $$W''(x) = [x(6 - \pi^2)W(x) + 2(3x^2 - 1)W'(x)]/[x(1 - x^2)]$$

for every point. In peak separation, a working vector of 17 points is used; therefore these computations must be done 17 times. Ideal peaks are generated for subtraction purposes and also for computing the quality measure of any residual peaks that turn up. It is clear that for any significant number of iterations, the amount of computation required quickly becomes prohibitive. Hence we recognized that a more efficient search method was imperative and did not pursue interval halving further.

2.3.2 Quadratic Approximation Searches

Since the quality measure, being computed from a sum of squares, is a quadratic form, it seems likely that it will also be approximately quadratic in any of the parameters, at least in a small neighborhood about the maximum. (In the case of the phase parameter, there could be difficulties, since its domain is circular; in practice, we found the approximation applicable.) Accordingly, a search process was tried in which the new value was defined as the vertex of a parabola passed through three previous values. That is, if a parameter value p_i resulted in a quality q_i, then the points (p_i, q_i), $i = 1, 2, 3$, define a parabola in the (p, q) plane; the axis of the parabola will be at

$$p' = \frac{1}{2} \cdot \frac{q_1(p_2{}^2 - p_3{}^2) + q_2(p_3{}^2 - p_1{}^2) + q_3(p_1{}^2 - p_2{}^2)}{q_1(p_2 - p_3) + q_2(p_3 - p_1) + q_3(p_1 - p_2)} \qquad (2\text{-}3)$$

This method did not always converge, because if the trial values were far from p', then small fluctuations in the q-values frequently caused disproportionately large errors in p'.

Next, we tried computing p' from a parabola fitted by least squares to four equally-spaced parameter values. The use of equally-spaced values offers some simplifications in computation. If $x_i = p_i - \bar{p}$, where $\bar{p}$ is the mean of the four values, then the new value will be

$$p' = \bar{p} - \frac{1}{2} \cdot \frac{\Sigma xq}{\Sigma x^2} \frac{(4\Sigma x^4 - (\Sigma x^2)^2)}{(4\Sigma x^2 q - \Sigma x^2 \Sigma q)}, \qquad (2\text{-}4)$$

where the sums are taken over $i = 1$ to 4. Additional protections are incorporated in the program to ensure (a) that the parabola is convex upward, so a maximum is found instead of a minimum, and (b) that the resulting p' lies between p_1 and p_4, so any remaining tendency to gross errors in p' will be curbed. If (a) fails, the p' is perturbed an equal amount in the direction away from the minimum given by (2-4); if (b) fails, then p' is forced to the p_i having the highest quality (typically p_1 or p_4); in either case, the search for a maximum is repeated until requirements (a) and (b) are both met, subject to the constraint that in no case shall the search be repeated beyond five iterations.

Once a maximum that satisfies requirements (a) and (b) has been found for a given parameter, the search is stopped and the process moves on to the next parameter; hence in most cases only 4 test values are required per parameter. In spite of this, the number of computations required to resolve an overlap is still very large, and application of the search method to routine separation was neither contemplated nor tried. Instead, it was applied only to overlaps on which the simpler method had failed--that is, to peaks whose resolved components were found not to fit properly into either talker's harmonic series. Furthermore, even among these peaks the process was restricted to peaks whose amplitude was large, on the grounds that small peaks would not degrade speech quality significantly if they were imperfectly resolved, while large peaks would. To judge by its performance on this small number of difficult overlaps, the process cannot be called an unqualified success; it was able to improve on the simpler program's performance in only approximately one third of all cases. It may be that we are up against an inherent resolvability limit in the other two thirds, however.

2.3.3 FM Effects

Efforts were also made to improve estimation of FM effects due to changing pitch. The occurrence of frequency shifts has the effect of introducing nonuniform phase across the harmonic peak. The reason for this nonuniformity can be seen in the approximation given in (2-2): changing frequency introduces a quadrature term whose amplitude is proportional to the FM rate and whose shape is the second derivative of W(x). When μ is zero, the phase across each peak is constant; when it is nonzero, the quadrature term has the effect of "cupping" the shape of the phase. This phase effect should be included in the ideal estimates in order to achieve accurate separation. In the basic separation method, however, it was impossible to do this, since we could not distinguish phase shifts due to FM effects from those due to the overlap itself. With the search method, the benefit of including FM rate as one of the parameters to be perturbed was uncertain, since the performance of the search method turned out to be mediocre in any case.

A different approach to the FM problem consisted of shortening the time window slightly. Since narrow spectrum peaks require wide time windows, and since the spectrum peaks must be narrow enough to handle the low end of the male pitch range (about 60 to 70 Hz) without excessive overlaps, there is not much we can do with window width, but it was shortened from 75.9 ms to 62.2 ms. Since the amplitude of the quadrature term in (2-4) is proportional to T^2, the benefit of this change is somewhat greater than one might expect: while the new window is about 5/6 the width of the old, the FM effects are about 2/3 of what they were. Nevertheless, it is obvious that this change is not going to work miracles for us.

Another question which has been raised by our experiences in this project is whether the importance of FM effects is as great as we have supposed. In practical terms, the only thing that counts is the quality (i.e., clarity and naturalness) of the recovered speech signal, and the evidence so far is that this is not greatly dependent on FM rates. In processed natural speech, the output quality is not materially worse during rapid pitch glides than it is over steady pitch, and in fact all of the defects that the current project is attacking were observed in the monotone samples processed at the end of the previous research effort. It may be that the best policy, at least in the short term, would be to shelve further work on FM problems until other problems, which have first-order effects on quality, have been solved.

3.0 PITCH EXTRACTION

The central problem of talker separation is the accurate determination of each talker's fundamental frequency. The separation process proceeds by selecting the spectral components which are the harmonics of the desired voice. The harmonics of one talker cannot be distinguished from those of the other without knowledge of both talkers' pitches. In the previous project, a pitch extractor was devised which was based on an extension of the Schroeder (1968) histogram. In this adaptation, a histogram was formed containing sub-multiples of the largest and most overlap-free harmonic peaks in the spectrum, and the largest entry in the histogram was taken to be one of the talkers' pitches. Then all harmonics of this pitch were identified and flagged, and a second histogram was formed using only unflagged peaks; the other talker's pitch was found from the largest entry in this second histogram.

The extractor worked spectacularly well on the vocalic speech data with which the first research effort was concerned, but failed badly with natural speech. The extractor made both octave errors (picking pitches an octave above or below the correct pitch) and random errors (the result of spurious histogram maxima caused by "unfortunate coincidences" among the harmonic peaks).

These errors had been encountered in vocalic speech as well, but they were then less frequent and had been brought under control by a number of ad-hoc fixes inserted at various points in the process, and also by coupling the pitch extractor to the pitch tracker. It did not appear that further fixes would be adequate to ruggedize the process for natural speech, and considering the irregularities of natural-speech pitch, we doubted whether a tracker could provide adequate guidance. What was needed was some way of validating a pitch decision by a preliminary test before committing the process to it. This was accordingly the first order of business in improving the pitch extractor.

3.1 Improvements in the Pitch Program

3.1.1 Pitch Validation

It was observed that a wrong pitch decision could usually be identified by examining the harmonic-peak-assignment results for the first few harmonics. In a good pitch value, the harmonics were well-represented, perticularly the low-order harmonics; in a bad value, they were poorly represented, and those that were present were often spurious values which fit the expected harmonic frequencies poorly. In errors an octave above the correct pitch, the missing harmonics were the odd-numbered ones; in other errors, harmonics were missing at random. In correct pitch values, the first five harmonics were virtually always present, except in the case of nasals and the "voice bar" of voiced stops. Hence the first validation test investigated was a trial peak-assignment run covering the first five harmonics.

In this test, a figure of merit was derived based on how closely the observed peaks fit the expected frequencies. A peak found to be right on target gets 5 points; this score diminishes linearly as the error increases, until a peak more than 10 Hz away (14 Hz in the case of the fundamental, which seems to fluctuate more than the other low-order harmonics) gets a score of 0. Missing harmonics score 0, and missing odd harmonics incur a 2.5-point penalty. The validity of the pitch estimate is the sum of these scores.

The measure just described is known as frequency validity. Two other validation tests were implemented in addition: the area of the corresponding histogram peak, and the sum of the heights of the peaks selected during the frequency-validity computation. The area of the histogram entry is an indicator of the number of spectrum peaks contributing to it and hence discriminates against entries created by random coincidences among spectrum peaks. The sum of the peak heights (which is normalized with respect to the square root of the total spectrum power) represents an attempt to favor those pitch values which account for most of the energy in the spectrum. Of these indicators, the frequency validity is by far the best discriminator and the peak-height test the least useful, possibly because it does not cover enough of the spectrum to comprise the bulk of the energy in it. We arrived at a combined validity figure by plotting scatter diagrams of validity results for a large number of frames. The total validity, v_t, is given by

$$v_t = .626\ v_f + .092\ v_a + .107\ v_h, \qquad (3\text{-}1)$$

where v_f is the frequency validity, v_a the amplitude validity, and v_h the histogram validity. Use of this validity measure will be described further below.

3.1.2 Multiple Histogram Entries

The success of the validity measure was great enough that we were able to abandon the double histogram process. Instead, a single histogram is used and all its maxima are evaluated (by means of the validation tests) as possible pitch values. The program compiles a list of the ten best candidates--i.e., those with the highest validities. This list of candidates is subjected to further screening (described below), and the two best surviving estimates are taken to be the desired pitches.

The use of a single histogram offers three advantages: (a) execution time is saved, since the program does not have to go through the process of forming and scanning a histogram a second time; (b) errors or uncertainties in the first pitch or in its harmonic peak assignments do not affect identification of the second pitch, as they did in the old process; and (c) knowing both pitches in advance permits certain improvements in the peak-assignment process.

The screening of the list of ten possible pitches consists of two steps. In the first, possible octave errors are detected and rejected; in the second, minimum thresholds are applied to reject spurious pitch values appearing in unvoiced speech.

The octave error test looks for possible harmonic relationships among the candidates. (For conciseness, we refer to octave errors, but any integer frequency ratio from 2:1 to 6:1 is investigated here.) If such a relationship is found, the higher-frequency estimate is rejected unless its validity exceeds that of the lower value by at least 3 points.

The voiced/unvoiced thresholds are defined as follows: first, the total validity must be at least 15 and the frequency validity alone at least 5; second, if the total validity is less than 21, then the frequency validity alone must be at least 11. If two or more pitches survive this screening, then both talkers are assumed to be phonating and the two highest-validity candidates are identified as the pitches; if only one value survives, then only one talker is phonating and it is up to the tracker to determine which one it is; if none survive, then neither talker is phonating.

A final test is used to avoid dual pitch values, which can result if the histogram entry for the correct pitch is bimodal as a result of submultiples being scattered among several bins. (This scatter can occur if one or more of the lower harmonic peak frequencies are badly perturbed by overlap.) In the absence of pitch crossings, therefore, once the first pitch has been selected, the second pitch is required to be at least 6 Hz away from the first.

3.1.3 Miscellaneous Changes

In addition to the above, histogram peaks are now based on the total area between minima, and the pitch estimate is computed from the centroid of the histogram peak; this is done to minimize the effect of scatter in the histogram. A new mapping from frequency to histogram index has also been introduced; it is of the form,

$$i = a + b/f, \qquad (3\text{-}2)$$

where a and b are scaling constants used to make the histogram cover the desired range and f is the frequency. This mapping has the advantage that the various submultiples of any given harmonic will be equally spaced along the histogram: for if a peak frequency is f_p, then the n^{th} submultiple will be mapped as $i = a + nb/f_p$, which is linear in n. By distributing the submultiples uniformly, we avoid any tendency for entries to pile up at one end or the other and bias the distribution.

3.2 Performance

The pitch extractor is now robust enough to handle natural speech with an accuracy ranging from 86 to 90 per cent. An example of performance with two male talkers is shown in Fig. 3-1. In (a) and (b), pitch values obtained for the individual voices are plotted; (c) shows the results for both talkers combined. In (c), x marks pitch values belonging to Talker 1 and o marks those belonging to Talker 2. As always with speech data, errors tend to cluster; in between clusters, the performance of the process is comparable to that of the old process on vocalic speech. Tracker coupling is not required and in fact has been dismantled, since it is not feasible in natural speech and would merely cause the errors in one routine to degrade the performance of the other. In heavy-error regions, the performance of the pitch process may still be unacceptable and may require intervention by the user--one reason why an interactive process would ultimately be desirable.

3.3 Alternate Pitch Extraction Methods

In addition to improving the existing pitch-extraction process, we also investigated two possible alternative pitch-extraction schemes. Among the principal modern methods for determining pitch, we have found the Schroeder histogram a particularly desirable starting-point, because it adapts readily to use with more than one talker and because it is a spectrum-peak-based method and hence naturally compatible with the rest of our basic process. Nevertheless, as we have seen, the histogram has its problems, and the question is always present, whether there may not be some other method which is also adaptable to multi-talker speech and whose problems might be more tractable than those of the histogram.

Two alternatives were briefly studied, and rejected, in the course of the present effort. These were extraction with the aid of linear prediction and maximum-likelihood pitch estimation.

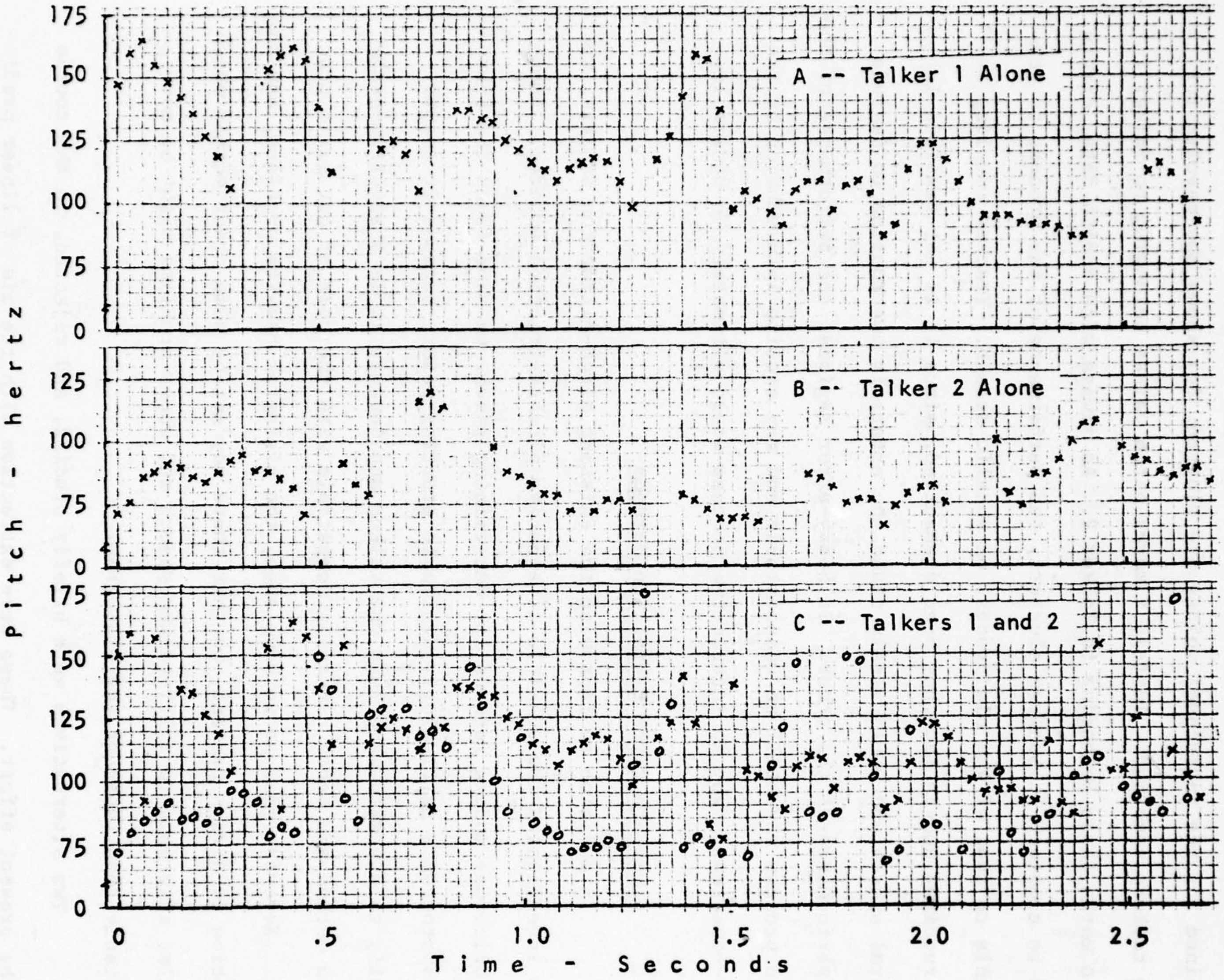

Fig. 3-1. Pitch Extractor Performance

Most linear-prediction pitch extractors operate on the principle of estimating and then removing the effects of the formants. The effect of the predictor on the speech signal is such that the prediction error tends to be a spectrally-flattened version of the speech waveform (Markel, 1972a). Pitch can then be found from this waveform either by autocorrelation methods (Markel, 1972b) or by taking advantage of the fact that each new glottal pulse is marked by an abrupt jump in the prediction error (Atal and Hanauer, 1971). Linear prediction derives its power principally from the fact that it embodies a particularly good model of the generation of single-talker speech. If the process is to be applied to two-talker speech, however, it must now somehow contrive to remove one set of formants from one component of the mixture and a different set of formants from the other. Obviously, this cannot be done, and although the operation might be approximated, the linear predictor extractor clearly admits of no straightforward generalization to two talkers as does the histogram.

The maximum-likelihood (ML) detector approximates voiced speech as a sum of periodic repetitions of a function $s_o(t)$:

$$s(t) \approx \sum_k s_o(t - kt_o).$$

(The approximation comes about because in real speech variations in pitch and formants cause t_o and s_o to change as functions of k.) Then the problem is to find t_o to minimize the mean-squared approximation error,

$$e = \int_0^T [s(t) - \sum_k s_o(t - kt_o)]^2 \, dt. \qquad (3\text{-}3)$$

Discussions of this process may be found in Wise et al (1976) and Friedman (1977).

The first point to be noted about this process is that if a value

t_o minimizes (3-3), then all integer multiples nt_o also minimize it. This is because the minimization process does not make any a priori assumptions about $s_o(t)$, and hence is not able to recognize the fact that if

$$\sum_k s_o(t - kt_o) = \sum_k s_o'(t - nkt_o),$$

it is only because $s_o'(t) = \sum_{r=0}^{n-1} s_o(t + rt_o)$. Hence we note that we are back to octave errors again.

Furthermore, we note that the ML detector is in fact a time-domain equivalent of the Schroeder histogram, as pointed out by Friedman (1977). The two approaches are linked by the elementary Fourier transform relation,

$$\sum_k s_o(t - kt_o) \leftrightarrow \sum_k S_o(f)\, \delta(f - kf_o), \quad f_o t_o = 1,$$

and indeed in Schroeder's process, the spectrum peaks are assumed to be a sum of integer multiples of an unknown fundamental f_o, while in the ML detector the time function is assumed to be the repetition of a signal at an unknown interval t_o. The practical significance of this for us is that if we were to substitute the ML detector for the Schroeder histogram, we would be substituting an equivalent method which has the added drawbacks that it works in the time domain (thus making no use of the peak tables) and that its extension to two-talker speech is not nearly as straightforward.

4.0 PEAK ASSIGNMENT

It was clear from the start of the present effort that many of the failings of the process when applied to natural speech were the result of errors in the peak-assignment phase of the pitch program. It was hoped that superior peak-separation accuracy and improved pitch estimates would enhance the performance of the peak assignment process, but this did not happen. It was clear that the process as designed was not able to reject spurious peaks originating with the other talker and was also not always able to recognize its own harmonics, particularly after a low-energy region between formants in which no peaks could be recovered.

4.1 Improvements in Peak-Assignment Techniques

To improve the assignment program, we investigated (a) new ways of computing the target frequency (i.e., the estimated frequency of the next harmonic to be assigned a peak), (b) ways of enabling the logic to reject spurious peaks, and (c) ways of bridging interformant gaps.

4.1.1 Target-Frequency Estimation

In the existing version of the process, each new target frequency was obtained by adding the estimated pitch to the old target. (This process is described in detail in RADC-TR-75-155.) If the search for the previous harmonic had found a peak which fit the target well (i.e., with an error of less than 3.5 Hz), then the peak frequency was averaged in with the old target before computing the new target. In this way, the assignment process stepped adaptively from one harmonic to the next. The purpose of this adaptive averaging was to protect the process against cumulative errors resulting from the uncertainty of the original pitch estimate.

In the present research, we investigated an alternative method for computing the target frequency. Here, on a successful match, the pitch estimate itself is updated, instead of the target. That is, for harmonic n, the new pitch estimate will be

$$p_{est}' = .8\ p_{est} + .2\ f_p/n, \qquad (4\text{-}1)$$

where f_p is the frequency of the assigned peak. We expected that by continually updating the pitch in this way, we could bring the cumulative errors within acceptable limits. Peak-assignment performance of this rule (known as adaptive multiplication, since the new target is obtained by multiplying the updated pitch estimate by the new harmonic number) was observed over a sequence of ten frames and compared with the old method (known as adaptive stepping). The errors in target frequencies when compared with known harmonics were as follows:

Method	Errors (Hz)		
	μ	σ	Maximum
Adaptive stepping	-1.86	6.45	14.88
Adaptive multiplication	-1.78	6.48	14.90

It was felt that the new method showed no clear-cut superiority over the old, and so the old method has been retained.

4.1.2 Protection Against Spurious Peaks

If we examine the spectrum of a pair of vowels, we will see regions in which the two talkers' harmonic peaks are interlaced, alternating with regions where they tend to overlap. Clearly, in the interlacing regions, the likelihood of selecting the wrong talker's peak is small, since the other talker's peaks are far away from the target frequency. In the overlap regions, on the other hand, the other talker's harmonic peaks will be very close to the target frequency and may be selected by mistake. In such a case, the adaptive stepping process, acting on the basis of a spurious peak, may introduce spurious "corrections" into the target frequency which will increase the likelihood of further incorrect assignments as the process begins to move out of the overlap region.

In view of this problem, we programmed the assignment logic to disable the averaging-in operation whenever it was in the vicinity of an overlap region. The locations of these regions were predicted as follows: let the two pitches be p_1 and p_2. The two trains of harmonics can be regarded as periodic signals of "quefrency" $1/p_1$ and $1/p_2$, respectively. Then the overlap regions correspond to the "beats" between these two signals, and from elementary considerations, the beat quefrency is $1/f_o = 1/p_1 + 1/p_2$; hence what we have called overlap regions will be at multiples of $f_o = p_1p_2/(p_1 - p_2)$.

Speech processed with this revised assignment logic showed no detectable abatement in crosstalk, however, and this process has been removed from the assignment logic.

4.1.3 Pre-Assignment

The existing assignment process fails to make use of one important source of information. The pitch estimate is obtained from a maximum in the frequency histogram, and this maximum, in turn, is composed of submultiples of a number of spectrum peaks. Hence these peaks are by definition harmonics of the pitch. That being the case, we start off with a set of known harmonics which the existing process then fails to use.

We therefore have prefaced the peak-assignment program with a program which finds and flags all of the harmonics that contributed to the histogram entry from which the pitch was found. (This process is known as "pre-assignment".) The regular assignment routine is now modified as follows: first, in normal operation, before searching for a new harmonic peak, the program sees whether a peak has been pre-assigned to that harmonic. If it has, then the pre-assigned peak is used, the search is skipped, and the next target is computed directly from the frequency of the pre-assigned peak (i.e., without averaging). Second, if more than two consecutive harmonics are missing, the program does not attempt to extrapolate past the gap in the usual way; instead, it looks ahead for the next pre-assigned peak and resumes the process from there.

These modifications offer several important advantages. First, execution time is saved by omitting the search procedure when it is not needed. Second, using known harmonics (where they are available) as a basis for subsequent targets provides much better protection against cumulative error than does the averaging method. Third, searching for a pre-assigned peak at an interformant gap provides an error-free way of bridging such gaps.

4.1.4 Supplementary Searches

It has been observed that a train of harmonics after an interformant gap may be completely missed by the existing process because of cumulative errors. The errors may not be so large as to prevent any of the peaks from being found at all, but they will be greater than 3.5 Hz and so will prevent the logic from locking onto the new peaks by means of the averaging rule. The purpose of the 3.5-Hz averaging limit is to prevent the assignment process from locking onto spurious peaks. Therefore, if this limit were widened with increasing frequency to permit the process to lock onto the higher harmonics more easily, one of our chief protections against crosstalk would be lost. Although pre-assignment provides the ideal solution to this problem, it depends on having conveniently located pre-assigned peaks, and since that in turn depends on random variations in the speech signals, it cannot be depended upon.

In the new process, if an interformant gap occurs which cannot be bridged by finding a pre-assigned peak, the process will now make a supplementary search for any sequence of peaks separated by the proper spacing. If a sequence of four or more such peaks is found, then these are accepted as harmonics, and the assignment process resumes at the beginning of this sequence. We have found that this is a satisfactory way of bridging gaps without sacrificing such immunity from spurious peaks as we now have.

4.2 Performance

On the basis of listening tests, the use of pre-assignment and supplementary searches has been found to enhance the performance of the peak assignment process greatly. The muffled sound is gone from the output speech, and there appears to be somewhat less crosstalk than with previous versions. The peak-assignment process still requires more work, however, since crosstalk and quality continue to be problems with natural speech. One possibility which should be explored in furure research is to enable the assignment process to work backwards (i.e., in the direction of decreasing frequency) when a pre-assigned peak has been found on the far side of an interformant gap, since we have no assurance that the peak which has been found is the first available one. Furthermore, the histogram formation procedure in the pitch extractor should be modified, if this can be done without increasing the pitch error rate, so as to increase the number of contributing peaks, particularly above 2.5 kHz. Ideally, the majority of harmonic peaks should be identified by pre-assignment, since this is both the most accurate and the most economical method available to us.

5.0 TRACKER ACQUISITION AND PREDICTION

Since the two-talker separation process proceeds by one 62-ms frame at a time, provision has to be made to ensure that the voice that was identified as (for example) Talker No. 1 on one frame will continue to be identified as Talker No. 1 on subsequent frames. As was explained in RADC-TR-75-155, the only practical test of identity in such short samples is continuity of pitch tracks. The pitch tracker matches the current pitch values to predictions made from past values and establishes talker identity on the basis of the match; it then predicts a new pair of pitch values for the next frame. In addition, it notes whether the pitch tracks are likely to intersect in the near future and makes this information available to other routines.

In the existing version, the tracker had two modes of operation, which determined how the prediction was to be made. *In the acquisition mode*, a zero-order predictor was used--that is, the next value was predicted to be the same as the current value. In the track mode, the next value was predicted by fitting a least-squares straight line to the three previous values and projecting this line to the next frame. (Four values were used in the vicinity of a pitch-track crossing.) A prediction was considered correct if it was no more than 5 Hz away from the actual pitch value; the tracker entered the track mode if two successive pitches matched the predictions and reverted to the acquisition mode if two consecutive errors occurred.

The zero-order predictor was too rigid to acquire many of the tracks that were encountered in natural speech, since a pitch track slewing faster than 5 Hz/frame could never be acquired. Although the straight-line predictor had given good results on vocalic-speech data, we felt that for best results a predictor based on the observed statistics of pitch tracks was to be preferred.

5.1 Improvements in the Tracker

5.1.1 Predictive Filter Design

The new predictor is designed using well-known statistical prediction techniques. The value of the pitch at frame n is estimated as a linear combination of the k values for the k preceding frames. That is,

$$\hat{p}(n) = \sum_{i=1}^{k} a_i p(n - i). \tag{5-1}$$

The predictor coefficients {a} which minimize the mean-square prediction error are computed from the covariance matrix of the observed pitch values. (These techniques are explained in detail in Makhoul (1976), although it is important to distinguish the present application, which is the prediction of pitch values, from what is commonly termed "linear prediction of speech".)

Since the pitch tracks available were of limited length, we had difficulty in gathering enough data to provide a statistically reliable basis for computing the predictor. In order to get as much data as possible, we used pitch tracks from the natural-intonation vocalic-speech data, which are long and unbroken, and we combined data for different tracks and talkers by averaging their covariance matrices.

Predictors were computed for k = 2, 3, 4, and 5. Predictor coefficients and prediction errors (standard deviations) are given in the following table, with the results from straight-line predictors provided for comparison. The performance of the statistically-derived predictors is clearly better.

TABLE 5-1

PERFORMANCE OF VARIOUS PITCH TRACKERS

		Prediction	% Errors
Order	Coefficients	Error (Hz)	> 5 Hz
2	(1.5994, -.6008)	3.13	9.65
3	(1.5432, -.4851, -.0599)	3.14	10.00
4	(1.5565, -.5168, .0349, -.0758)	3.09	10.26
5	(1.5082, -.4106, .0499, -.2165, .0687)	3.01	9.21
SL(3)*	(1.3333, .3333, -.6667)	3.72	15.63
SL(4)*	(1.0000, .5000, 0.0, -.5000)	4.01	16.03

*SL(n) = least-squares straight line over n frames.

5.1.2 Acquisition and Prediction Rules

In the new tracker, we define a track break as any jump of more than 10 Hz. On the first frame after a track break, we use a zero-order predictor, as before, but on the k^{th} frame after a break, we use a (k - 1) order predictor, whether the tracker is in acquisition or track mode.* Thus the tracker is in a sense trying to track all the time, and the 10-Hz rule means that tracks can be acquired even when slewing at a rate of 10 Hz/frame. The tracker still requires two consecutive correct predictions to enter the track mode, and in the track mode, a discrepancy of 5 Hz or more is considered an

*The highest order predictor currently used is order 3; we do not feel that higher-order predictors are advisable until they have been computed from samples of natural speech rather than vocalic speech.

error. The likelihood of an error can be related to the derivatives of the actual pitch values (see Appendix); for a third-order predictor, for example, there will be no error if

$$|.3933\ p'(x) + .4682\ p''(x)| < 5\text{Hz},$$

where $p(x)$ is the pitch at frame x. With a frame-to-frame spacing of 31.1 ms, these limits translate to a maximum first derivative of 409 Hz/s and a maximum second derivative of 11,030 Hz/s^2. The straight-line predictor, on the other hand, had no limit on $p'(x)$, but made an error if $p''(x)$ exceeded 2.33 Hz/frame^2, or 2,210 Hz/s^2. Empirically, the likelihood of such errors can be estimated from the last column of Table 5-1, which shows the percentage of frames in the data for which the prediction error exceeded 5 Hz.

5.2 Tracker Performance

A sample of the tracker output is shown in Fig. 5-1. Pitch values, as corrected by operator intervention, are plotted as points, and the predictions made from these points are plotted as broken lines. Pitch values belonging to Talker 1 are indicated by x and those belonging to Talker 2 by o. The lines joining the predictions are shown dotted when the tracker is in the acquisition mode and solid when in the track mode. Notice that each individual predictor spends relatively large amounts of time in the acquisition mode, but that, in spite of this, pitches are assigned correctly most of the time. This is because (a) the filters attempt to predict regardless of the mode they are in and (b) as long as at least one prediction is correct, the matching rule will assign the talkers correctly. Tracks are occasionally inverted, as at the points marked A and B, but these events are infrequent enough that they can be corrected by user intervention. We

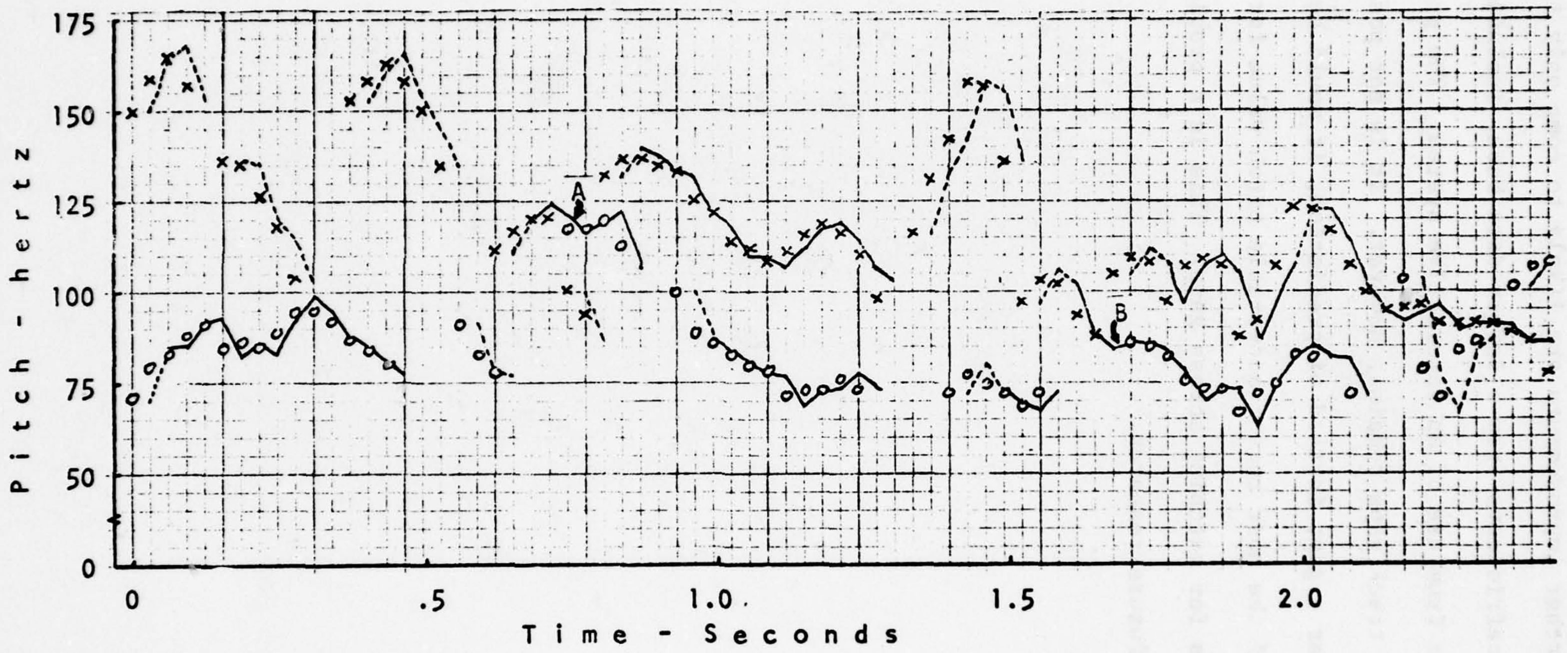

Fig. 5-1. Pitch Tracker Performance

believe that further improvements are likely to come only from recomputing the predictor coefficients from a larger data base, including natural-speech pitch tracks, or from use of an interactive system. The problem of effectively bridging track gaps remains, however. It is not possible to predict pitch values over a gap; instead of prediction, it would be best to maintain a record of the most recent acceptable pitch value for each track and use these values for matching across gaps. This is a problem which should be attacked in future research.

6.0 PROCESSING OF NON-VOCALIC SPEECH

In extending the original process from vowels and vowel-like sounds to fully-natural speech, one of the chief questions to be answered is what to do about nonvocalic speech sounds. The process is designed on the assumption that its input will be a sum of periodic functions and its output a single periodic function; accordingly, the process of synthesizing the output function does nothing but insert harmonics into the transform. For our purposes, therefore, nonvocalic sounds are speech sounds that are not periodic, such as plosives and unvoiced fricatives, or that have significant aperiodic components, such as voiced fricatives. There is no provision in the existing program for inserting such aperiodic components into the output transform.

It would not be difficult to modify the synthesis program to insert noise into the transform as needed, and there is reason to believe that the exact amount and type of noise to be added would not be critical. Cherry and Wiley (1967) reported that speech from which non-vocalic sounds had been gated out regained its intelligibility significantly when wideband noise was inserted in all the silent intervals. Apparently the brain is used to wide variations in the sounds of fricatives and plosives and so will accept any suitably-placed noise burst as the required sound. The chief problem in handling these other sounds is detecting their occurrence, since noise does not admit of being separated by techniques such as those used in our process. In planning for this phase of the research, we expected to use the voiced/unvoiced indicator, which would be a by-product of pitch validation, as our guide. We could then follow the example of Cherry and Wiley and insert a small amount of noise into the transform in the absence of voicing. In practice, the voiced/unvoiced indication does not appear to be functioning very reliably. We identified voiced and unvoiced passages in each individual talker's voice by processing the voices separately, observing the pitch values obtained, and checking them by examining the peak tables. Using these results as a guide, and comparing the performance of the process on two-talker speech, we saw that on many unvoiced frames, the pitch program gave a spurious pitch value rather than an unvoiced indication. (Unvoiced frames are identified by a pitch of 0.) The thresholds described in Section 3.1.2 are used for the voiced/unvoiced decision; if they are raised in order to reject these spurious values, we find that they will also reject many voiced frames as well.

Another important result came out of the runs with single-talker data, however, because we were able to play back the output speech generated on these runs. Here we had the process working on single-talker data without interference and synthesizing only what the pitch extractor gave it to synthesize--i.e., only the periodic portions of the speech. On playback, the speech was clear and realistic, and the ear perceived substantially all of the nonvocalics that had been present in the input speech. This is an important observation, because it indicates that actually very little need be done for handling nonvocalics, and in particular, that faking by means of Cherry and Wiley's method may not be required. (Part of the reason for this may be that in the experiment by Cherry and Wiley, only the loudest portions of the original speech were not gated out, which probably meant that many of the formant transitions, which play a crucial role in consonant perception, were lost. The two-talker process seems to preserve formant transitions with great fidelity.) The single-talker experiments also suggest that the main thrust of future research should probably be to improve the quality of the output provided by the existing process, so that the output with two-talker data will be comparable to that with single-talker data, rather than to seek special ways of handling non-vocalic speech.

APPENDIX

Effect of Pitch Variations on Tracker Accuracy

Given a predictive tracker of the sort described in Section 5.1.1, we wish to determine the effect of slope and curvature of pitch tracks on the tracking error, and in particular to know for what values of p'(x) and p"(x) the tracker will make errors greater than 5 Hz.

The predictor produces the following estimate of the forthcoming pitch value:

$$\hat{p}(x) = \sum_{i=1}^{k} a_i p(x - i). \tag{A-1}$$

We will use a Taylor series expansion of p(x) and assume all derivatives beyond the second negligible; then

$$p(x - i) = p(x) - ip'(x) + \frac{1}{2}i^2p''(x), \tag{A-2}$$

and the error is

$$\begin{aligned} e &= \hat{p}(x) - p(x) \\ &= \sum_{i=0}^{k} a_i[p(x) - ip'(x) + \frac{1}{2}i^2\ p''(x)], \end{aligned} \tag{A-3}$$

where $a_0 = -1$. Substituting the predictor coefficients derived for the pitch tracker and ignoring the p(x) term (which is negligible), we get the results tabulated on the following page.

k	Error
2	$-.3978\ p'(x) - .4019\ p''(x)$
3	$-.3933\ p'(x) - .4682\ p''(x)$
4	$-.3244\ p'(x) - .7047\ p''(x)$
5	$-.3142\ p'(x) - .7158\ p''(x)$
SL(3)	$-2.3333\ p''(x)$
SL(4)	$-2.75\ p''(x)$

SL(n) refers to least-squares straight-line predictors computed over n frames.

Note that the statistically-derived predictors show a linear trade-off between $p'(x)$ and $p''(x)$. Since we wish to limit errors to 5 Hz, we can translate these expressions into hertz and seconds, assuming an interval of 31.11 ms between frames. In that case, we get the following limits on $p'(t)$ and $p''(t)$:

k	Max $p'(t)$ (Hz/s)	Max $p''(t)$ (Hz/s^2)
2	404	12 850
3	409	11 030
4	495	7 330
5	512	7 220
SL(3)	∞	2 210
SL(4)	∞	1 180

REFERENCES

B. S. Atal and S. L. Hanauer (1971), "Speech analysis and synthesis by linear prediction of the speech wave", J. Acous. Soc. Am., v. 50, no. 2 (part 2), pp 637-655.

E. C. Cherry and R. Wiley (1967), "Speech communication in very noisy environments", Nature, v. 214, p. 1164.

D. H. Friedman (1977), "Pseudo-maximum-likelihood speech pitch extraction," IEEE Trans. Acous., Speech, and Sig. Processing, v. ASSP-25, no. 3, pp 213-221.

J. Makhoul (1975), "Linear prediction: a tutorial review", Proc. IEEE, v. 63, no. 4, pp 561-580.

J. D. Markel (1972a), "Digital inverse filtering--a new tool for formant trajectory estimation", IEEE Trans Audio and Electroacous., v. AU-20, no. 1, pp 129-137.

J. D. Markel (1972b), "The SIFT algorithm for fundamental frequency estimation", IEEE Trans. Audio and Electroacous., v. AU-20, no. 5, pp 367-377.

M. R. Schroeder (1968), "Period histogram and product spectrum: new methods for fundamental-frequency measurement," J. Acous. Soc. Am., v. 43, no. 4, pp 829-834.

J. D. Wise et al (1976), "Maximum-likelihood pitch estimation," IEEE Trans. Acous., Speech, and Sig. Processing, v. ASSP-24, no. 5, pp 418-423.

MISSION
of
Rome Air Development Center

RADC plans and conducts research, exploratory and advanced development programs in command, control, and communications (C^3) activities, and in the C^3 areas of information sciences and intelligence. The principal technical mission areas are communications, electromagnetic guidance and control, surveillance of ground and aerospace objects, intelligence data collection and handling, information system technology, ionospheric propagation, solid state sciences, microwave physics and electronic reliability, maintainability and compatibility.

AMERICAN REVOLUTION BICENTENNIAL
1776-1976